Article

New Tools for Old Tasks: A New Approach to the Investigation of Malay

Asian Linguistic Anthropology
2020, Vol. 2(3) 21-38
(c) The JALA 2020
Article reuse guidelines
Jala.pub/permissions
DOI: 10.47298/jala.v2-i3-a2
Jala.pub

Zuraidah Mohd Don
Universiti Teknologi Malaysia, Malaysia

Gerry Knowles
Independent Scholar

Abstract

Digital tools designed for linguistic analysis offer new ways of approaching old tasks, and they now routinely enable tasks which were formerly impossible. The work for this paper is based on MaLex, which is a collection of data tables and procedures designed to represent the intuitive knowledge of speakers of Malay, and provides the infrastructure for the solution of problems in linguistic analysis. This paper reports the use of the MaLex parser to investigate the adjectival system of Malay, including superlatives and the formation of manner adverbials. Although linguists have always been able to identify possible syntactic rules and try them out on small datasets, the automatic parser is able to extract examples from a large corpus, and it is much more effective than a linguist in ascertaining the ordering of rules, and tracing their interaction. Since the examples in this paper are also syntactic constituents, they are the appropriate units for translation, and in this case they are translated into English. The phonological component concatenates the phonological representations of constituents to form higher level structures, and an extension to Malex which is planned but not yet completed is intended to increase the range of waveform annotations that can be used as input for linguistic analysis.

Malay is a suitable language for this research, because although it is under-investigated in relation to its importance as one of the main languages of ASEAN, it has extensive written records which make it possible to compile large corpora for research. A human linguist can get started on a very small amount of data, and the same is true of the approach pioneered by MaLex. For this reason, MaLex could prove to be a suitable model for the digital investigation of insufficiently researched languages.

Introduction

The United Nations declared 2019 The International Year of Indigenous Languages. The associated conferences and other events that took place must have prompted many linguists not directly involved in the study of indigenous languages, including the present authors, to consider what could be done in practice to promote the study of these languages, and where possible protect them from extinction.

It is generally acknowledged that the world's languages are dying out at such a rate that many will become extinct without being recorded adequately or even at all, and that what will be left to posterity is an inadequate and biased sample of the languages that survived into the modern period. Since the resources available for the study of indigenous languages are insufficient in relation to the size of the problem, new ways must be found to preserve as much as possible before endangered languages eventually become extinct.

Ways need to be found to maximise the impact of the individual researcher. At present, only a small proportion of the linguistic knowledge accumulated by a researcher in the course of a lifetime's work ever finds its way to publication. Major reviews such as the work of Benjamin (2012) and Blust (2013) on the languages of ASEAN are based on decades of work, of which only the parts deemed worthy of publication have been preserved, and most of which has been irretrievably lost. We would know a lot more if we had access to linguistic knowledge that generations of fieldworkers thought not worth publishing.

A related problem is that it is difficult to continue where some researcher in the past has left off. Taking Hendon (1966) as an arbitrary example, it would take a considerable time to ascertain the contribution to knowledge made by Hendon at sufficient depth to incorporate it in a new project. It is surely much easier to start afresh on some new variety. As a result, Hendon's work has little if any relevance for researchers after just two generations, and its value is likely to be lost.

The point made at some length in this paper is that the means have long been available to solve these problems and others by the appropriate use of digital resources. The essential step is to convert data into computer-readable form. For example, different researchers use different systems of phonemic representation, but as long as the systems are reasonably consistent, they can be normalised to some standard system. There are many ways of representing morphology, including the use of plus signs and round brackets; but again, these can be normalised, and in fact morphology can be represented much more clearly using digital storage.

MaLex ("MALay LEXicon") is a collection of data tables and procedures designed to represent the structure of Malay in digital form. The work that has gone into its development is the kind of work practical linguists have always done, and includes ascertaining the phoneme system and other aspects of the phonology, working out how the morphology works, and identifying syntactic rules. However, MaLex goes beyond traditional practical linguistics by implementing procedures developed within corpus linguistics, including stemming, tagging and parsing. MaLex is thus not just a static model, but has procedures that put theoretical ideas into practice in the automatic or automated analysis of naturally produced Malay texts.

MaLex was originally modelled on EngLex, which is a companion collection of data tables and procedures for English, compiled from the word lists and procedures developed for different projects. It came as a surprise just how much of the design of EngLex could be re-used with modifications for Malay. A model that can be used for languages as different as English and Malay can also be used for many other languages. For closely related languages, the possibilities are even greater. Much as the data from different researchers can be normalised, the data for one language can in some cases be converted into data for a related language.

Once linguistic data is encoded in digital form, it can be manipulated in ways that were formerly impossible. An important design feature of MaLex is that linguistic information is in principle never thrown away, but a place is found for lit somewhere in a table or a procedure. Much of this information corresponds to the linguistic knowledge possessed by researchers but not deemed worth publishing. Needless to say, automatic procedures draw all the time on information that might otherwise be regarded as useless. A significant property of MaLex is that it can be re-used. Tables and procedures can be exported and modified for use in any number of projects, and the Malay lexicon can be used as a digital dictionary. In this way, MaLex can be developed to increase the output of a research team working on indigenous languages.

The next section, The Theoretical Foundations of MaLex, describes MaLex in some detail, including how it evolved, and its theoretical foundations. The following section, What MaLex Can Do, shows how it can be used in its present state to carry out useful tasks in linguistic analysis. The Discussion section explains how MaLex could be extended to expedite the analysis of related languages, and this leads into the conclusion.

The Theoretical Foundations of MaLex

The design of MaLex is based on the premise that language has a real structure that exists independently of any theory, and that the purpose of linguistic theory – surely like any scientific theory – is to identify the patterns that explain the data. In accordance with this approach, it is the linguist's task to identify and describe the patterns in the data.

Old Tasks, New Tools

The manner in which practical linguists go about their work depends at least in part on the technologies available and applied. Until about fifty years ago, much traditional linguistic analysis applied no special research technology at all apart perhaps from the printing press used to publish research findings. For this reason, linguists have traditionally carried complex data sets around in their heads, together with procedures used to process the data. Philologists retain in their memories lists of reconstructed lexical items and the sound changes that apply to them. Practical linguists memorise huge numbers of examples of linguistic structures of many different kinds, together with associated analytical procedures.

Although linguistic analysis might seem to be a discipline entirely *sui generis*, unrelated to what researchers do in other disciplines, the measures taken to solve problems are in fact special cases of more general research techniques. Much of what practical linguists do in the analysis of languages involves the application of procedures to data sets. Greatly improved general computer-based solutions have been available for about forty years. The modular and procedural Pascal language appeared in the 1970s, and *inter alia* enabled automatic linguistic procedures. The manipulation of complex data sets was facilitated by the introduction of relational databases, beginning with Ashton-Tate's dBASE. Even before these technologies became available, the scope of linguistic analysis was extended by the application of digital methods, beginning with the work of Henry Kučera and Nelson Francis (1967), to tackle data sets far too large to process in the traditional manner. In the present century, there is no good reason for linguists to use their own memory rather than computer memory, or to use their own brains as a central processing unit.

MaLex has from the beginning used digital storage for data sets, and formal procedures written in a computer language to manipulate them, and it has been used to tag millions of words of naturally produced corpus material. To that extent the research paradigm to which it belongs is clearly corpus linguistics. However, it differs from conventional corpus linguistics in that the aim is to use text processing to create a language model for use in the performance of useful tasks. Practical linguists have always created language models by analysing language systems such as the phonology, the morphology and the syntax; but these models have largely resided inside their heads. MaLex presents the whole of the model explicitly in the form of data tables and formal procedures.

Digital technology not only makes it possible to present the language model in a manner formerly impossible, but also resolves a long-standing disagreement. Since the 1960s (Chomsky, 1965), the notion has been current that *contra* practical linguists, the linguist's task is not to account for data but to model the intuitive knowledge possessed by the native speakers of a language. Although linguistic analysis itself is by nature difficult, there is no need for the output of the analysis to be difficult to understand. For native speakers of Malay, the MaLex data tables bring into conscious awareness linguistic information that their unconscious brains are using all the time, and for that reason the content of the tables can be and should be obvious. Practical linguistic analysis and modelling native speaker knowledge are complementary to each other, and there is no necessary theoretical gulf between them.

The Development of Speech and Language Technologies

The work of linguists has for the last fifty years or so had an ambivalent relationship with the development of speech and language technologies. For example, grammatical tagging is the first step in processing corpus texts, and tagged texts are typically used as input to technologies. On the other hand, the technologies are created not by linguists but by computer scientists and engineers ("speech scientists").

Fred Householder (1952) made an important distinction between what he called "God's Truth" and "hocus pocus" linguistics. God's Truth linguists deal with entities that really do exist in some form in the data they analyse, and this approach is taken by MaLex, traditional practical linguistics, and mainstream corpus linguistics. Hocus pocus linguistics is understood in different ways but includes the view that language consists of a mass of formless data, and that the researcher's task is to impose some order and structure upon it. This is the dominant approach in the development of speech and language technologies. The aim in this case is not to understand language but to produce a working product. Problems of language structure are either ignored or avoided by the use of "black box" techniques, such as neural nets and Hidden Markov Models.

There has been growing interest in recent years among speech scientists in recycling resources developed for languages like English for the benefit of less well-resourced languages. Malay is less well-resourced than English, and can benefit from resources developed for English, and can be used in turn as a source for other Austronesian languages. While some linguists might deprecate black boxes, the reality is that they make it possible to provide such resources as speech synthesis and automatic speech recognition for languages that might otherwise have no resources at all.

One reason speech scientists resort to hocus pocus methods is that nothing better is readily available, even for Malay. Asmah (1993) presents a cornucopia of valuable information concerning Malay morphology and syntax; but the linguistic description needs to be presented in a form that can be used in speech technologies. There are several published accounts of Malay phonology that present it in accordance with some current theory; but speech scientists cannot be expected to read PhD theses and other publications in order to find out which if any of them can be used in the development of speech technologies. Practical linguists may grumble at the quality of the linguistic analysis used for Malay speech technologies, including the use for Malay of phoneme symbols designed for English; but it seems nobody's responsibility to bridge the gap between practical linguistics and speech technologies. If a speech scientist were to set out to produce a speech synthesis system for some little-known Austronesian language, it would be difficult to know where to start. This is where the approach pioneered by MaLex comes in. MaLex can be re-worked for some other language, and a derivative of MaLex would provide the tables and procedures required as input to speech technologies.

The MaLex approach is the tortoise in relation to the black box hare: its methods take much longer than black box methods, but they extend linguistic theory and understanding into areas for which black boxes provide solutions without explanations. The view taken here is that black boxes are valuable and indeed necessary for the development of speech and language technologies in areas for which conventional solutions are not available, but that when the tortoise eventually catches up, the solutions made available have the potential to lead to improved technologies.

The Presentation of Theoretical Insights

The last century saw the formulation of a large number of competing linguistic theories presented on the printed page for publication. In many cases, these are not easily transferred to digital form, and it is necessary to separate out the theoretical insight incorporated in a theory from the manner in which it happens to be presented. Competing theories that incorporate different insights make different falsifiable claims at least one of which can in principle be falsified, whereas those that differ in presentation make more or less the same claims but in different ways. The best illustrative example is perhaps the branching syntactic tree, which is a presentation of a fundamental insight into syntax, but which is often taken to constitute the insight itself. In this case, the development of the parser was held up for a long time until a means was found to represent syntactic structures in tables.

The transfer of insights into digital form has important potential consequences for linguistic theory. This is because once an insight is incorporated into a working module, it can be difficult or impossible to say which particular theoretical presentation it comes from, because it more probably distils the insight from several different sources. For example, the operation MERGE is generally associated with the minimalist program (Chomsky 1993), but a remarkably similar idea presented in a different way has been taught for centuries to beginning learners of Latin. It is impossible to say whether the corresponding idea in the MaLex parser comes from modern linguistics or traditional language teaching. Grammatical transformations are generally regarded as a twentieth century insight (Chomsky, 1957, 1965), but they were routinely taught in the Latin class. Linguists with experience of analysing real language data can expect to find the presentation of insights novel and unfamiliar; but the insights themselves should be familiar enough.

In view of the incompatibility of the manners in which different theories are presented, it is not possible to pick some arbitrary collection of linguistic theories – such as X-bar syntax, natural phonology, and componential analysis – and expect them to work together. The theoretical insights themselves have to be sound in their own right, and it must be possible to re-present them in digital form so that they work together like the parts of a Rolls-Royce engine. This means that there are tight constraints on the linguistic theories that can be incorporated into the design of MaLex.

The Case of the Disappearing Phonemes

Undue attention to presentation can obscure the nature of the insight itself, and this is illustrated by the conventional representation of pronunciation. The lexical entry in MaLex for *kaki* 'foot, leg' includes the pronunciation /kaki/. These symbols do double duty, both as unique identifiers for Malay phonemes, and to represent the modal phonetic value of these phonemes. In the context of moveable-type technology, the use of complex symbols to represent different kinds of information simultaneously is economical and intuitively easy to learn and understand. However, it creates problems when transferred to digital technology. In MaLex, the symbol "k" is connected to a phonemes table in which it is associated with values in three separate fields that identify it as voiceless, velar, and as a stop. Now if the phonetic value is encoded in three fields, it does not have to be encoded again in the symbol "k" itself. The essential property of a unique identifier is that it is unique, and the characters used to represent it are of no theoretical significance whatsoever. In fact, "k" could be replaced by any character that the database engine can handle. Although we might think of phonemes as real components of language, and even speculate where they might be stored in the brain, the logic of the relational database indicates that they are actually artefacts of a medieval technology.

Half a century ago, there was a long debate whether phonemes or features were the units of the spoken language, and it might seem that phonemes have to yield to features. However, if we relate features to acoustic events in speech waveforms, the same applies again. Allowance being made for some simplification, "voiceless" doubles as an identifier and descriptor for the absence of voicing, "velar" the same for formant transitions into and away from the vowel, and "stop" the same for closure and a release burst. Like phonemes, features have no existence independently of their roles as links between levels of linguistic description.

The labels used for features are useful mnemonics, but they are of no theoretical significance, and "velar" could equally well be called "George". This is important, because although Malay /k/ is described as a voiceless velar stop, it actually occurs as a glottal stop before consonants and in final position. In this case, the logic of the relational database clears up a long-standing Lilliputian controversy. Phonemic transcription implies a sequence of acoustic events in the corresponding waveform, and is an efficient means of representing sequences of articulatory and acoustic events.

The Evolution of MaLex

The MaLex project began with a request from Dewan Bahasa dan Pustaka, the Malaysian government body responsible for supporting the Malay language, to produce an automatic grammatical tagger for Malay texts. Since it was known that as in the case of many other languages the grammatical class of Malay complex words can generally be inferred from the morphology, the first collaborative task was to develop a stemmer to identify affixes and isolate the simplex form. The output from the stemmer was used as input to the tagger, and in this way MaLex from the first integrated theoretical insights from different sources.

Linguists working on languages they do not know well come across all sorts of information that must belong somewhere in the linguistic description, although exactly where it belongs is so far unknown. In the case of MaLex, as a matter of principle, snippets of information were stored in tables. For example, the information that the word *membaca* 'read' consists of the prefix *mem-* attached to the stem *baca* was stored in a table that eventually became the main lexical table. The form *mem-* is a variant of the prefix spelt *meng-*, and this information was stored in a related table accessed by the stemmer. Simplex forms are in general suitable names for lemmas, and *baca* was stored in a related lemmas table. The information in the tables is obvious to anyone who knows the language, and lexicographers will recognise the connection between lemmas and headwords. Information of this kind must have been stored many times in fieldworkers' notebooks and later discarded as trivial when the language was better understood; but in the case of a computer table, many examples of trivial information constitute a valuable resource.

It is not difficult to work out how to analyse a word like *membaca*, and the steps taken to analyse this and other words were recorded in a procedure. Linguists unfamiliar with programming might find it difficult to understand the procedure itself written in a computer language, but the sequence of logical steps that it encodes can be explained in a manner that can be readily understood. In a sense, although the manner of presentation may be unfamiliar, everything in the database is obvious. Of course, *ars est celare artem*: simplicity is not achieved immediately or without effort, and requires a huge amount of work.

Although MaLex began with a stemmer and tagger, it soon became clear that in order to carry out useful tasks it would have to increase its functionality. When working on a language one does not know well, it is difficult to keep track of the meaning of the items one is analysing, and so for the benefit of the second author, it was necessary to develop a means of indicating the meaning of words. Since the meaning of complex words is related to their morphology and so to their grammatical class, this was not an arbitrary unrelated extension, but an extension building on what was already available.

In some cases, the solution of one problem suggested a related problem requiring investigation. For example, the prefix *meng*- has several variants which may seem arbitrary both in their form and their distribution, but which are easily explained by the phonological rules that govern allomorphic variation in prefixes. The allomorphic variation provides the key to understanding initial mutation, e.g. in forms such as *menulis* 'write', in which *meng*- is prefixed not to **nulis* but to the stem *tulis*. As in the case of the Celtic languages, initial mutation makes the stemmer more complicated but provides a theoretically motivated explanation for patterns that must otherwise seem quite bizarre. These rules were eventually combined with a spelling-to-phoneme algorithm to add a pronunciation component. The addition of a phonological component makes the language model more complex, but enables it to account for a large additional amount of data, and to undertake a wider range of tasks. Phonological information was used, for example, to generate phonetic specifications including F0 and durations for speech synthesis, using written sentences as input. The more patterns are investigated and understood, the more the complexity of the data is reduced to order.

The big problem was parsing, partly because so many competing approaches to syntax have been put forward since the 1950s that it was difficult to decide which route to take. However, it is a commonplace notion in computer programming that complex problems can be solved by dividing them into simpler modules, and dividing progressively until the parts of the problem are sufficiently simple to code directly. Some of the research problems to which MaLex was applied involved searching the corpus data for patterns of words, and these were solved by this same process. Eventually the search procedure was developed to make a (partial) syntactic parse of the corpus data. Whereas conventional parse trees require some linguistic expertise to understand, the MaLex parses are included in tables, and are obvious to anyone who knows the language, e.g. that the phrase *ke dalam* 'into' contains *ke* 'towards' and *dalam* 'inside, deep'. A parser with glosses is already half way to translation, and it is currently being developed into a component to translate Malay texts into English.

What MaLex Can Do

MaLex contains the linguistic infrastructure required to carry out useful tasks, including investigations of the language itself which would otherwise be difficult or impossible. The purpose of this section is to give some indication of what MaLex can do now at its present stage of development. The information reported here is retrieved by means of a concordance procedure with the ability to process a corpus to extract trigrams consisting of a key word or key grammatical class and the preceding and following words, e.g. all sequences of three words with a passive verb in the middle. The data is taken from two corpora consisting of articles selected from the Malay newspaper *Berita Harian*. The first is currently being compiled, and it now contains about 330,000 words, this figure being increased by the daily addition of new selected articles. It is used for preliminary testing, and records linguistic items that have occurred at least once, but includes no frequency information. Part of this corpus has been used in the development of the parser, and so it contains some parsing information. The other is a fixed corpus which is intended to yield quantitative information. A subset of over 500,000 words has been tagged, and this subset has been searched by the concordance procedure, and instances counted.

The following three searches were undertaken in the context of investigating the grammar of kata sifat 'adjectives'.

Superlative Adjectives

Malay is like English in having two means of forming superlative forms, one using an affix (the English suffix -*est* and the Malay prefix *ter*-), and the other using a separate word (English *most*, Malay *paling*). Malay *ter*- also has an unconnected use as a verbal prefix, and the corresponding data has been ignored here.

The starting point is to consider what had already been recorded in all the corpus data so far processed, *ter*- forms being listed in the main lexical table, and *paling* forms listed in a syntactic table. When the investigation of adjectives began, 178 superlatives had been processed, 66 formed only with *ter*-, 78 formed only with *paling*, and 34 having both forms. There is no clear connection with meaning, although words of English origin such as *strategik* and *konservatif* form superlatives with *paling*, while common adjectives such as *baik* 'good' and *baru* 'new' take *ter*-, often (as in these two cases) alongside *paling* forms.

In the subcorpus searched by the concordance procedure, there were 1179 *ter*- forms in total, but just 38 different forms. This procedure counts instances, and provides a frequency distribution. Just ten words had *ter*- forms with a frequency of 10 or more: utama 'main' (381); baik (251); baru (103); tinggi 'high' (99); akhir 'last' (88); besar 'big' (77); dahulu 'before' (51); dekat 'near' (42); buruk 'ugly, bad' (17), penting 'important' (10). There were 16 words with frequencies between 2 and 9, and 12 singletons.

There were also 243 *paling* forms in total, and 93 different forms. There were just five with a frequency of 10 or more: penting 'important' (16); tinggi 'high' (11); teruk 'serious' (11); cergas 'active' (11); popular (10). There were 36 with frequencies between 2 and 9, and 52 singletons. These distributions give an interesting insight into this area of Malay grammar. The use of *ter*- is concentrated on a small group of frequent forms, all of which are frequent words anyway. Superlatives for other words can be formed with *ter*-, but these are not frequent. *Paling* forms are less frequent overall, and they are more evenly spread over different forms. No one form is particularly frequent in comparison to *ter*- forms, and few occur more than ten times in the data.

The raw figures also conceal a number of warning signs. Although superlatives may be associated with adjectives, the superlative forms retrieved are not all adjectives. *Utama* and *dahulu* are not adjectives, and the status of *dekat* is ambiguous. *Paling popular* may be consistent with the accepted means of forming superlatives with English adjectives, but whether it counts as a well-formed phrase of Malay may be subject to disagreement.

Manner Adverbials

There are three main ways in which Malay manner adverbials are formed:

1. *Cara* or *secara* (both here glossed 'manner') are followed by an adjective
2. This combination is optionally preceded by *dengan* (by default 'with')
3. *Dengan* is followed immediately by an adjective, without *cara* or *secara*.

In this case, some preliminary information is available from a small set of parsed texts. A total of 349 different adverbials are recorded, beginning with cara (29), secara (171); and dengan (149). *Dengan* is followed by cara (9); secara (3); an adjective (129); or by miscellaneous items (8).

In the data retrieved by the concordance procedure, *cara* (212) occurs 24 times followed by an adjective: *terbaik* 'best' (11) and *lain* 'other' (4), the remaining cases being singletons. *Secara* (773) is a much more frequent word, and occurs 291 times before an adjective. In fact, *secara* is followed in principle not just by a simple adjective but by an adjectival phrase; and it is followed by words such as tidak 'not' (48); lebih 'more' (19), and complex adjectives. A trigram concordance is not powerful enough to deal with these cases, which need to be analysed after a parse has been completed. Just a few simple adjectives follow *secara* at least 10 times: langsung 'direct' (40); haram 'religiously forbidden' (20); rasmi 'official' (17); aman 'calm' (12); automatik (10); peribadi 'personal' (10) , sah 'legally valid' (10).

Dengan (4426) is a high frequency word, and it occurs 303 times before 86 different adjectives, only 7 of which occur at least 10 times: baik (45); penuh 'full' (24); segera 'quick' (19); teliti 'careful' (14); mudah 'easy' (14); sempurna 'perfect' (14); jelas 'clear' (12). The most frequent type is clearly type 1, but with *secara* rather than *cara*, and followed by type 3. Type 2, in which *dengan* is followed by a phrase beginning *(se)cara* does occur, but it is not very frequent. A historical linguist will surely infer that *dengan (se)cara + adjective* was probably the original type, and that it was generally simplified to either the noun phrase *(se)cara + adjective* or the rather unusual construction *dengan + adjective*.

Ber<noun> + Adjective

The prefix *ber*- has a number of uses, and in many cases marks an intransitive verb, e.g. *bergerak* 'move', *bernafas* 'breathe', *berfikir* 'think'. It has a quite different function when attached to a noun, in which case it has a number of possible glosses, including 'with' and 'have', e.g. *berleher* 'with a neck', *berbahasa* 'having language'. In the case of clothing, 'wearing' is a more appropriate gloss, e.g. *gadis bertopi besar* 'a girl wearing a big hat'. Although *ber*- is classed as a prefix, the stem can be modified by an adjective like an independent noun, e.g. *binatang berleher panjang* 'animal with a long neck', *orang berbahasa Melayu* 'person with Malay'. Whereas the combination of *ber*- and noun are rather awkward as words, the addition of an adjective creates a phrase that more comfortably matches the notion of a syntactic constituent.

Perhaps a better analysis is to treat *ber*- in this case as a clitic with a role like that of a preposition. The constituents of the phrase *berbahasa Melayu* are not the two orthographic words, but *ber*- and (*bahasa Melayu*). The problem is that as a bound form, *ber*- cannot be written as a separate word according to the rules of the writing system, and for that reason it has to be attached to the noun. This is not a unique or otherwise unknown situation: as long ago as 1762, Robert Lowth observed that in the phrase *the King of England's soldiers*, *'s* is attached to *England* but applies to the whole phrase *the King of England* (Lowth, 1762).

The parsed data yielded only 31 different types, e.g. *berteknologi tinggi* 'high tech', *berkelajuan tinggi* 'high speed', which look like regular collocations. In two interesting cases, the noun acted as a classifier of the adjective, namely *berwarna hitam* 'black in colour', *bersaiz kecil* 'small in size'. Otherwise, the sample was too small to give any clear indication of what kinds of noun and adjective are involved in this construction.

The concordance procedure identified 7087 instances of *ber <noun>*, including 609 instances in which the next word was an adjective. These 609 instances include 129 different adjectives, of which just seven occur at least ten times: tinggi 'high' (63); tempatan 'local' (33); baik 'good' (25); sementara 'temporary' (16); besar 'big' (16); rendah 'low' (12); antarabangsa 'international' (11). Forty six adjectives occurred between 2 and 9 times, and there were 76 singletons.

In view of the hint in the parsed data that ber- may be attached to regular collocations, it was decided to check combinations of noun and adjective. Among the nouns combining with *tinggi* 'high' were mutu 'value' (11); teknologi (10); wajaran 'justification' (10). *Tempatan* 'local' followed *kuasa* in 32 cases, and *baik* 'good' followed *nasib* 'luck' in 18 cases, while *semantara* 'temporary' followed *berkenaan* 'about' (8) and *sifat* 'characteristic' (4), *rendah* 'low' followed *pendapatan* 'income' (10), and there was one instance of *sekolah rendah* 'elementary school', which is a known collocation. *Antarabangsa* 'international' followed *taraf* 'status' (9).

It was expected that ber<noun> would also be followed by numbers, e.g. *beristeri dua* 'having two wives'. Numbers were indeed found, not as a count of items but as indices, e.g. of ages following *berusia* and *berumur* ‘aged’, and of amounts following *berjumlah* ‘in total’.

Discussion: The Potential Uses of MaLex

When a new technology is introduced into research practices, the initial response is to do what was done before but more rapidly, more efficiently, or on a bigger scale. MaLex likewise began by carrying out the kinds of task long undertaken by practical linguists, supplemented by more recent tasks developed within corpus linguistics. However, the technology brings to light hidden problems, and offers new opportunities. It is difficult to construct a parser without considering what syntax is all about, and this line of thinking led to the distinction between insights and presentation. The design of the phonemes table led almost by accident to the concept of disappearing phonemes. In the early stages of MaLex, the different components – stemmer, tagger, parser, spelling-to-phoneme algorithm – were written separately *ad hoc* in accordance with the conventional understanding of levels of language; but because they are all closely connected in the context of the language, the result was an unacceptable degree of overlap in code and tables: The components have been progressively re-written so that they fit into an integrated system: improvements made to one component bring to light ways in which other components can be better integrated. As a result, MaLex is becoming leaner and more fit for the purpose of solving linguistic problems, not just for Malay but also potentially for other languages.

Work on Malay

The immediate applications of MaLex are in the study of Malay, which is after all what it was designed for. Since it is not dedicated to the solution of any specific problem or set of problems, it can be used to solve a wide range of problems, because the general infrastructure is already in place to support the necessary extensions. For example, the future development of the parser depends on an intelligent concordance procedure to identify the data to be analysed, and in this case it was possible to extend the existing functionality to make the analyses presented in the section above.

The concordance procedure made it possible to investigate the structure of Malay at a deeper level. The formation of superlatives is taken to be a matter of free variation: Sneddon (1996, p. 180) states that Malay superlatives are formed with *paling* or *ter*-, and Asmah (1993, p. 153) states in relation to *ter*- that Makna yang didukungnya ialah “paling” ‘the meaning which is conveyed by it is “most”’. Although *paling* and *ter*- can indeed be treated as free alternatives at the level of the syntactic rule and associated meanings, they clearly differ in their patterns of collocation. It could be that Malay is on the way to forming a syntactic rule that *ter*- occurs in fixed collocations with a small set of frequent adjectives, whereas *paling* is used generally for superlatives formed on the fly.

The case of manner adverbs is rather different, because a set of three options seems to be in the process of reducing to two. There may be a tendency for some adjectives to occur with one rule or the other, but the evidence so far available is insufficient to show a pattern, and the frequency figures do not indicate any pattern at all. In the case of *ber*- attached to a noun, there are some indications that the noun and associated adjective in some cases form a collocation, but again the evidence is insufficient to draw any firm conclusion. Accounting for the intuitive knowledge of native speakers involves rather more than bare syntactic rules.

Using MaLex for Other Languages

If EngLex can be used as the model for MaLex despite the distance of the relationship between English and Malay, then presumably MaLex can be used as the model for the analysis of closely related languages. There is no need for researchers to start at the beginning when investigating a new language. Input data tables and procedures for Malay can be modified as required for the new language in order to create output tables the contents of which are obvious to native speakers.

It has been known since the publication of Grimm's Law (Grimm, 1822) that the key to understanding relationships among historically related languages is the identification of cognates, i.e. linguistic items which correspond in form and meaning, and which can be presumed to have been inherited from some common source and not borrowed from some third language. Philologists have conventionally built up a mental stock of cognates, e.g. German *Bein* 'leg' and English *bone*, or German *Dach* 'roof' and English *thatch*, and used them to infer the sound changes believed to explain the derivation of the extant linguistic forms from the common source.

Using MaLex, it is possible to reverse this process. Starting with a small word list such as a Swadesh list, it is possible to identify cognates and infer the rules that relate their forms to those of the first language. It is of course not possible to predict with certainly that the cognate will exist at all in the second language, but if it does exist, its form can be predicted. For example, given the standard Malay forms *tiga* 'three', *hijau* 'green' and *udang* 'prawn' and selecting Kelantanese as the second language, it is predictable that if these words have cognates at all in Kelantanese, their forms will be respectively [tiɡɔ], [iʤa], and [udɛ]. More generally, it is possible to devise an algorithm to derive Kelantanese forms from their cognates in standard Malay.

What applies to phonology applies elsewhere in the formal structure of the language. Cognates are likely to be found in the morphology, and MaLex morphology could be re-used with the necessary modifications. Syntactic rules would apply to significantly different data, but if correctly formulated they might well apply with little modification. Of course, different varieties are likely to have undergone different semantic changes, as in the case of *Dach* and *thatch*; but this is not a serious problem in Germanic philology, and judging by available word lists, it would not appear to be a serious problem in Austronesian linguistics either.

The groundwork undertaken in the development of MaLex can in principle be used to the advantage of researchers investigating related languages to make much more rapid progress than would be possible using conventional means. Researchers do not have to spend their time investigating properties of the language which are actually predictable, and can concentrate more profitably on interesting unpredictable linguistic phenomena.

The advantage of using algorithms to derive cognates is that they represent real properties of the related languages. Kelantanese speakers really do say [idʒa] where speakers of standard Malay say [hidʒau], and the same applies to ‘three’ and ‘prawn’. It is a truism in historical linguistics that each succeeding generation copies the language of the preceding generation, but with some modifications. The identification and reversal of modifications enables the linguist to go back in apparent linguistic time.

Historical Relationships among Languages

The establishment of relationships among languages and varieties is close to historical reconstruction. An important concern in ASEAN linguistics is to reconstruct Proto-Austronesian and the historical ancestor of the Austroasiatic languages to parallel the reconstruction of Proto-Indo-European. Although reconstruction on this scale is beyond the scope of this paper, MaLex nevertheless has the potential to make a significant contribution.

The reconstruction of forms believed to have belonged to an earlier and unrecorded stage of the language is necessarily a hocus-pocus exercise. Protolanguages and the intermediate stages through which they allegedly passed on their way to modern languages are also hocus pocus constructs. Reconstructions are of great value because they impose order on sometimes huge amounts of linguistic variation spread over time and space, and enable linguists to comprehend the data in an organised manner, and exchange historical information with each other. However, reconstructions do not recapitulate history. Tracing history is an exercise of an altogether different kind, and typically requires knowledge of emporia and other centres of power[1]. For example, although it is legitimate to construct algorithms to predict the cognates of standard Malay forms found in related varieties such as Kelantanese, Iban and Kerinci, it would be illegitimate, not to say absurd, to infer that present-day standard Malay is the historical ancestor of the other varieties[2].

The origins of Malay belong somewhere in the history and geography of the region south of the Kra isthmus, and to trace the history we would need to know much more than has survived about centres of power including Sri Vijaya, Melaka and also the Kingdom of Patani (Ibrahim Syukri, 1985). The maps showing the geographical distribution of cognate forms published by Asmah (1977, pp. 32–37) give an insight into real linguistic history not obtainable from any other source. An extension to include distributions in west Borneo would make it possible to test the claim made by Adelaar (1992, pp. 206–207) that the Malayic languages including Malay first emerged from west Borneo.

MaLex contains no historical information, but it could be extended to deal with languages and dialects – what Adelaar (1992, p. 1) calls "isolects" – encoded for geographical location. The focus would not be on the internal structure of the isolects themselves, but on the geographical distribution of linguistic forms[3]. Linguistic change does not take place in conveniently discrete languages and dialects, but is the outcome of competition between linguistic forms, some of which will be cognates and others unrelated lexical items. The items under consideration should not belong to a restricted set such as a Swadesh list, but include all relevant forms for which distributional information is available. Swadesh lists are of cultural value, and the Leipzig-Jakarta list (Haspelmath & Tadmor, 2009) usefully excludes loan words, but short lists are generally unsuitable for the task of reconstruction.

For the traditional philologist, reconstructing a dead language or language family was a major undertaking, and perhaps for that reason reconstructions have been regarded as faithful accounts of how later languages reached their contemporary form. From a computational point of view, a reconstruction is just another algorithm. There are many possible reconstruction algorithms, and their output depends on the input assumptions and the languages selected for the reconstruction[4]. It is to be expected therefore that different scholars will produce different reconstructions of Proto-Austronesian, and different accounts of the intermediate stages through which languages allegedly passed as they evolved into their modern forms. These differences are embarrassing if they are believed to be competing versions of historical events, but there is no problem at all if they are recognised as the outcome of different algorithms.

Conclusion

This paper has made the case for the use of digital technologies in the solution of linguistic problems, and for their potential use in expediting the study of indigenous languages. The essential characteristic of MaLex is that it has taken the leap from traditional to digital methods, and records linguistic information in computer-readable form. The technologies themselves have long been established in computer science, and they have the advantage that they closely approximate the way linguists have traditionally thought about language structure and solving linguistic problems. In this way, the MaLex approach minimises the leap from old to new. Digital technologies have developed rapidly in recent decades, and procedural languages have long been dismissed as old fashioned by some computer scientists, and questions have long been asked concerning what kind of database management systems will eventually replace relational databases. These are not serious problems, because tables and procedures can always be updated in accordance with later developments.

A second important characteristic of MaLex is that it can be reproduced *mutatis mutandis* for other languages. A rapid start can be made on a new language by using MaLex as a model, identifying the mutanda, and generating a new model with modifications. The particular niche suggested here is the exploitation of cognates. Generated models with location information can generate geographical distributions to strengthen efforts to ascertain synchronic relationships among languages and dialects, and make reconstructions more geographically valid, and perhaps fit them into history. If this can be achieved, MaLex will have made a positive if belated response in connection with the International Year of Indigenous Languages.

References

Adelaar, K. A. (1992). *Proto Malayic*. Canberra: The Australian National University.

Asmah, Hj. O. (1977). *The Phonological Diversity of the Malay Dialects*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Asmah, Hj. O. (1993). *Nahu Melayu Mutakhir (4th ed.)*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Asmah, Hj. O. (1995). *Rekonstruksi Fonologi Bahasa Melayu Induk*. Kuala Lumpur: Dewan Bahasa dan Pustaka.

Benjamin, G. (2012). The Aslian Languages of Malaysia and Thailand: An Assessment. In McGill, S. & Austin, P. K. (eds.), *Language Documentation and Description* (Vol. 11, pp. 136–230). London: SOAS.

Blust, R. (2013). *The Austronesian Languages* (Revised edition). Canberra: Australian National University.

Chomsky, N. (1957). Syntactic Structures. Berlin: Mouton.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge: MIT Press.

Chomsky, N. (1993). A Minimalist Program for Linguistic Theory. *MIT Working Papers in Linguistics*. Cambridge: MIT Press.

Grimm, J. (1822). *Deutsche Grammatik (2nd ed.)*. Berlin: Dieterich.

Haspelmath, M. & Tadmor, U. (Eds.). (2009). *Loanwords in the World's Languages: A Comparative Handbook*. Berlin: Mouton de Gruyter.

Hendon, R. S. (1966). *The Phonology and Morphology of Ulu Muar Malay*. New Haven: Yale University.

Householder, F. W. (1952). Review of "Methods in Structural Linguistics" by Zellig Harris. *International Journal of American Linguistics* 18, 260–268.

Ibrahim, S. (1985). *History of the Malay Kingdom of Patani*. Cleveland: Ohio University Press.

Knowles, G. & Mohd Don, Z. (in press). Reconstructing language and regional identities in early England: A critical perspective. McFarland.

Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present-day American* English. Providence: Brown University Press.

Lowth, R. (1762). *A Short Introduction to English Grammar*. London: A Millar, R & J Dodsley.

Sneddon, J. M. (1996). *Indonesian: A Comprehensive Grammar*. Global: Routledge.

Endnotes

[1] The best known reconstructed language family is probably Germanic, but one has only to read Caesar and Tacitus to realise that it does not fit into history.

[2] But see Asmah (1995). This is a genuine conundrum.

[3] The study of the distribution of regional dialect forms in England has yielded historical information far more detailed than has been possible using traditional philological techniques (Knowles & Mohd Don, in press).

[4] The discussion in Benjamin (2012, pp. 145–151) also raises the question what is represented in a family tree.